Table of Contents

5 Machine Consciousness

This chapter is a basic introduction to machine consciousness, a field which may eventually generalize the study of consciousness and sentience. This chapter's scope is restricted to only the coverage needed to understand and relate to the safety and security of strong AI and is not intended to be a comprehensive guide to the construction of cognitive architectures. This brevity is a form of focus, as there is a state of confusion that exists both within and around this subject; machine consciousness rests at the intersection between philosophy of mind, cognitive science, and computer science, and, like the field of strong AI it will one day enable, is in a state of constant flux. This is complicated not just by the complex underpinnings that presuppose the abstractions and problems of its conceptual space, but its unavoidable collision with theology and the false dichotomies in human knowledge between philosophy and science.

5.1 Role in Strong AI

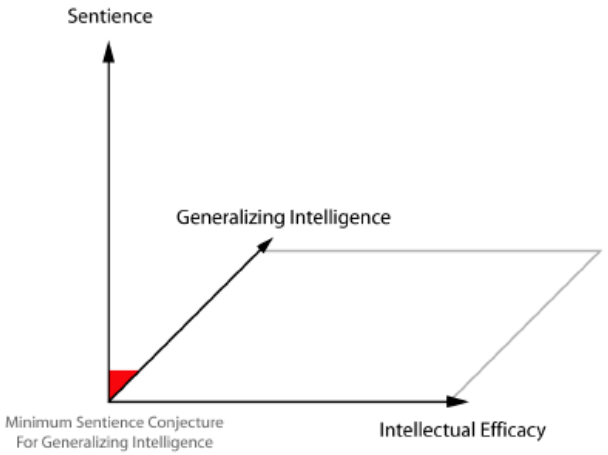
As defined in previous chapters, strong AI is within the set of all cognitive architectures. More clearly, this means that strong AI *must* be capable of undergoing experience as per its definition. This is not so much a claim about artificial intelligence as it is a stipulation of this particular class of implementations.

It is possible to create narrow AI implementations that are reasonably optimal, such that no strong AI could make a significant improvement upon them. For example, a machine code description, tuned by an expert, that enumerates the digits of pi, would be a narrow AI implementation of optimized pi enumeration. It is unlikely that even a highly intelligent strong AI would be capable of enumerating pi more efficiently, nor would it likely be able to *appreciably and significantly* optimize it beyond the micro-optimization done by a trained expert in programming at that level of description languages and on that particular architecture. This is because there is a distinction between generalizing intelligence and the optimization processes implied by non-generalizing forms of "intelligence", such as narrow AI. This distinction is powered by the claim that such a generalizing capacity is *not possible* without sentience. Thus, the assertion is that machine consciousness is a prerequisite for building strong AI.

Overlooked is the fact that strong AI will have a very wide range of intellectual capacities. We tend to focus solely on the beneficial (or harmful) extremes of this technology, with exclusion to minimal implementations. This is crucial to understand, as it makes a connection between sentience in general and that of strong AI. There are many organisms that would rightly be considered equal to strong AI, even though they do not *ostensibly* present the same intellectual capacity as some humans. Such a failure to recognize their intelligence is entirely on our part, in that we lack the ability to understand the inner world of such entities [1, 2, 3, 4, 5] sufficiently to make a true judgment as to their level of cognition, especially considering our inherent biases towards certain ends and aims in human cognition, e.g. the various trends in what constitutes the signaling of wit and status. This is obviously controversial, as this spectrum of generalizing intelligence is not commonly thought to depend upon sentience, and, by extension, apply to all sentient animals, but this is the exact claim being made. It leads to the stronger claim that sentience represents an *evolutionary advantage*

in that it would have been impossible to achieve generalizing intelligence without it. Stated directly, the claim is that, *at a minimum*, all vertebrates, and some invertebrates, possess some level of generalizing intelligence, and that this attribute is dependent upon their sentience, and that the ability to undergo experience presupposes the conceptual and analogical generalization faculties of generalizing intelligence.

Figure 5.1: The orthogonality between sentience, generalizing intelligence, and intellectual efficacy.



Sentience and intelligence are independent. Non-sentient processes can demonstrate efficacy at intellectual tasks. Sentient processes can lack any notion of intelligence, despite being able to undergo experience. Generalizing intelligence and intellectual efficacy are also independent. A generalizing intelligence may be capable of new intellectual endeavors but lack corresponding efficacy.

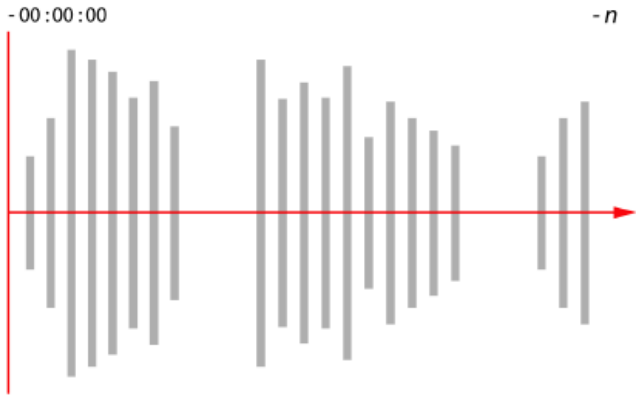
The figure is illustrating that sentience and intelligence are related but orthogonal. An entity may be sentient but have low or almost no "intelligence" by various standards. The opposite is also true: there are processes that achieve intelligent results that are non-sentient. This demands an explanation to which the answer begins with a question: to what is the *strong* in strong AI referring? The answer is that it is the property of having a mind, one that can undergo experience, grasp value, and understand meaning. This measure of *strength* is not in terms of the fidelity of the cognitive range of intellectual capacity but in the means to generate and endure the phenomena of experience. It is said to endure because it has no option of eliminating experience altogether while remaining extant.

The creation of a sentient artificial process will be trivial compared to the problem of implementing a generalized learning algorithm. It will be difficult to achieve consensus on the actual space and range of what constitutes experience, even among those who agree that it is possible to create. The greatest obstacle, however, will be in the disbelief and hostility that will arise towards the notion that such processes are sentient at all. This will be in spite of the fact that it will be falsifiable as to whether a particular process is *capable* of undergoing experience. However, being falsifiable does not mean that we will have solved all of the philosophical problems, which, may or may not ever receive a satisfactory answer. It is also possible that most of the philosophical questions will be explained away, in that they will no longer be considered valid questions. Despite this potential, the philosophy must not be dismissed out of hand, as it will be used to argue for and against important concepts that will underwrite ethics, jurisprudence, and politics, many issues of which are already beginning to be discussed in the mainstream.

Thus, the role of machine consciousness is not to create generalizing intelligence but to *enable* it. Sentience is necessary but not sufficient for generalizing intelligence; it does not directly address the sapient aspects of a cognitive architecture. That is to say, there exist possible strong AI implementations which are sentient but lack any appreciable intellectual quality that would enable us to communicate or relate to them. These *simplex* implementations would not even be aware that they are conscious, as they would lack any reflexive or meta-cognitive ability. Complex thought would not arise at this base level. The entity would exist in a purely *neutral* state of mind in which it would accept any experience that arose without resistance. This does not, however, include the *absence* of experience, as any sentient process is necessarily a real-time system.

To understand the real-time aspect of sentience we can resort to a simple analogy. Below is a constructed digital waveform representation:

Figure 5.2: A time-domain representation of the experiential stream of a sentient simplex.



The regions of zero signal or intensity depict not an absence of experience but the experience of absence. Nothingness is given concretion by the real-time requirement of sentient processing. The simplex must undergo a neutral fragment to experience "silence".

The gaps in the waveform represent what we would perceive as silence. If this were a representation of the experiential data of a *simplex* sentience (this being its one and only dimension of experience) the zeroed pulses would correspond with *neutral* states of mind. There is still an auditory experience, it's just unique in that it fills time with the least possible sound. To understand further, we must assert that *silence* and the *absence of the experience of sound* are two very different things. The gaps between the positive or negative pulses fill the conceptual space in the experiential encoding. To not have them would represent a discontinuity in this stream of experience, which would only be possible if the entity's consciousness were interrupted, paused, or stopped. Interestingly enough, a simplex would be fundamentally incapable of recognizing that it had been interrupted if the data stream were resumed without skipping information, but this would not work in a situation where the real-time system is perceiving its environment. Despite experiencing a discontinuity, however, a simplex sentience still wouldn't comprehend its significance. The term *neutral* is useful here, as a gap in the stream of experience is not an absence of experience but a neutral state as per the context of that stream. If still perplexed, consider the explicit rests in musical notation as an example. The presence and function of neutral states are essential to any stream of experience.

What this tells us about strong AI is that they will all need to be real-time

systems with respect to their streams of experience. This is perhaps one of the easiest ways to distinguish them from a typical narrow AI implementation, which has no such concepts or constructs as fundamental parts of its make up. Even if such a system ends up being real-time due to some application constraint, it does so at a level which is much further downstream to the processing than sentient calculation. This real-time demand for sentient processing sets up a minimum condition which corresponds with the notion that the subject must *endure* and *undergo* experience. This fuses any such implementation description with time; it must process to progress through steps to be realized at all, as it doesn't make sense as a static object or an intrinsic physical property. These are processes which involve the exchange and interpretation of information.

This hints at some of the trouble with viewing the reduction of conscious states to only the physical properties of things or the arrangement of their physical structure. All the confusion in the history of philosophy of mind hinges on the lack of acceptance of processes, with even some modern philosophers rejecting the effect of interpreters by throwing them out with abstract objects. In computation and interpretation, we're dealing with extents in time, which mean that the properties of things take on additional meaning through the semantics of their arrangement and interpretation in that extended dimension, and this is above and beyond their intrinsic physical properties. It is, of course, true that they are constrained in their arrangement and composition by their physical properties, but as long as sufficient states can be derived then we have the means to derive new properties through the semantics of the interpreter. This is neither metaphysical nor mystical. It's simply a matter of fact that things can be so arranged through time, with a corresponding interpreter, such that it gives rise to new functionality and new properties that are not present in the *static* representation of the underlying units of composition. This is routine in the spatial extents of alphabets giving rise to descriptions, such as books, images, and other data; at no time is there some Platonic potential [6] in some alphabet that gives rise to some data description. Rather, it is that through extension the object is realized and constructed by interpreting the arrangement of that alphabet co-extensive to the dimension in which it is projected. That the dimension is called *time* is of no special significance except with regard to the fact that we can only appreciably apprehend it in the moment. In storage, the time-like extents are afforded through the same spatial extents of arranging descriptions. The difference is the way in which that extent is interpreted. A string of characters representing someone's name has no time-like interpretation to be understood to be a name, but a film, or a sound, can only be manifested and apprehended by conscious entities in any given instant of ongoing experience; slowing, pausing, or interrupting it fundamentally alters the experience of it.

The real-time constraint for time-like extensions of objects is an assertion on the identity of the experience which must be taken to be unalterable if to be perceived as entailed by its description, e.g. watching a film below the intended rate at which it was encoded can be an instantaneously frustrating experience, and with good reason. Such a disparity between the experience of a time-like object and its description could be considered *noise*, and can be measured and quantified explicitly as the difference between the rate of processing and the intended rate implied or specified by its encoding (description language). This applies both to the interpreting process and the rate of perception in total; the rate at which something is playing back may be incomprehensibly fast or slow for the rate of potential perception. The reverse also applies to the rate of sentient processing.

What this all leads to is the question of the physical reduction of consciousness [7]. Indeed, the whole *does* equal the sum of the parts where the sum includes the time-like extents of the proper arrangement of those parts. This *will* lead to a *how* answer to the philosophical question of experience arising out of non-conscious physical states, but it doesn't necessarily explain *why* such a thing is possible at all. The answer to which

is perhaps too simple to be accepted: it is a tautological result of the interplay between an interpreter and the process it realizes; it makes it so.

Lastly, and perhaps most importantly, is the role that machine consciousness has in realizing *value*. Critically, without sentience there can be no ability to realize or apprehend value, and without a concept or ability to grasp value there can be no general moral or emotional intelligence. Thus, another major role for machine consciousness is to make *general moral intelligence* possible for artificial intelligence. This is distinct from simply applying moral efficacy externally by interpreting the behavior of a system as having consequences or agency or choices. This is because, unless a system has the capacity to derive and understand value, it is devoid of the experiential knowledge of the processes it caries out. In these cases, as it pertains to strong AI, we're referring to the ability to even *attempt* to reason about the morality of a decision, as distinct from its *efficacy* at general moral reasoning, which has parallels with the dichotomy between narrow AI and the generalizing abilities of strong AI. The relationship being that narrow AI may be able to instrument narrow moral intelligence, but would lack general moral intelligence for the same reasons it is fundamentally incapable of realizing generalizing intelligence. This is because *value* presupposes moral intelligence, which would be semantically meaningless without sentience to substantiate it [8]. This is true even if one could denote infinite non-conscious rule processing for which decisions were to be made; enumeration and rule-following in the absence of the capacity to reflect upon the process are not forms of moral intelligence, even if the particular rules result in what would be considered reasonable moral efficacy per some context. This is also why it is not even wrong to argue for moral intelligence as a means of safety, which, to date, has been exclusively implied to be the non-general and non-conscious moral frameworks of decision theory and micro-economic thought. These methodologies are fundamentally incapable of entailing the value that presupposes the reflective capacity required for moral agency. Naturally, the next question should be: then what exactly *is* value?

Definition: Value. *The experience of a positive, negative, or neutral sensation that accompanies or is associated with one or more experiences, with experience being that which is inclusive of all mental content, including, but not limited to, thought, knowledge, and perception. Value is further distinguished by being either intrinsic or acquired, with intrinsic value being a static association or accompaniment to one or more experiences, ab initio, by way of the underlying implementation, e.g. a pleasure-pain axis. This contrasts with acquired value, which is dynamic, capable of change, and is associated with acquired mental content, e.g. belief, knowledge, and actions.*

This definition appears to be endlessly recursive but is curtailed in practice. There can only be a finite set of experiences that can be instantiated for any given implementation of machine consciousness, and, further, the phenomena of experiencing something *as* intrinsically positive, negative, or neutral is terminated or rooted in experience itself, despite the appearance that it would continue to refer to other experiences ad infinitum. That is to say, value is experiential, but is one level of complexity or organization above it, and must not be confused with its referents, including knowledge and the beliefs associated with values, as there are certain types of value which are integral to the semantics of an implementation. For example, the intrinsic value of positive and negative sensations that surround the informational content of pleasure and pain as distinguished from the beliefs one has about these experiences.

This last issue, as exemplified by pleasure and pain, is subtle, as our unified stream of consciousness makes it confusing as to the separability of our experiences; we must recognize that, especially as it pertains to value, that our experience of that which it is associated with and the accompanying value that arises with it are *composite*. This is evidenced in humans through the clinical cases of pain asymbolia [9, 10, 11, 12, 13, 93], a neurological

condition in which the informational content of pain is disassociated or unaccompanied by the *intrinsic value* that normally follows [14]; those with this condition are capable of describing the intensity and quality of the pain, as if it were merely words being read off a page, but do not experience the negative value sensation that comes along with it. As a result, they have to form knowledge about this mental content and respond accordingly. This is non-trivial, as injury or death may result in the absence of the immediate and unconditional somatic and psychological urges of these intrinsic values. That is to say, they may realize it is negative, but, without the automatic and involuntary experience of intrinsic value, they have to rely upon the *acquired values* associated with the knowledge of the injury, which may not trigger the same priorities in salience, and may have increased delays in processing.

5.2 Sentience, Experience, and Qualia

In the last section, sentient simplexes were used to illustrate a hypothetical single dimension of experience. More complex and realistic cases of machine consciousness will require a complex mixture of multiple streams and types of experience. In the philosophy of mind, this is referred to as *binding* [15, 16, 17, 18, 19, 20, 21]. Such a subject is said to be *unitary*, in that the individual streams of experience have been combined into a *unified* and composite experience. For an analogy, think of the individual tracks of audio and video that are layered to produce a film, all of which are combined in such a way so as to allow a synchronous interpretation between its sights and sounds. When played back, it (hopefully) appears as a coherent and unified experience. Note that the usage of the word *stream* applies to both a quantized continuous or discrete interpretation in this text. This vernacular is borrowed from input-output (I/O) programming constructs, in which *discrete* units of information are read or written using buffers [22, 23]; this offers a counter-example where the stream terminology is used for a non-continuous source.

In this book, all streams of experience are described as being made up of *fragments*, which are referred to as *qualia* in the philosophy of mind [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37]. Qualia are literally the nature of "what it's like" to experience a fragment in a stream of experience. Examples of qualia include all sensations, sights, sounds, pleasure, pain, and even emotions and thoughts, depending on the philosophy. In this analysis all such fragments of experience should be considered qualia, and all contents of any possible mind are to be considered experience, including (non-exclusively to other aspects of mind) thoughts, memories, and knowledge [38, 39]. Qualia must not be confused with the *informational representation* of experiential fragments; there is a distinction between the knowledge of the fragment and the experience of the fragment. For example, one could provide the triple for red (255, 0, 0) but the overwhelming majority of readers will not experience that triplet as having a color different from the surrounding text unless they are synesthetic [40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51]. This is perhaps one of the more difficult notions to unpack from the philosophy of mind, as it requires an acceptance that such descriptions, despite being physically reducible, correspond to fragments of experience only insofar as they can be realized by the processes which enable sentience. That is to say, *the correlation between these fragments and what the subject is experiencing is a product of the sentient process and is not innate to the descriptions themselves*, which are merely used to invoke the semantics of the implementation.

Straight away, the two most difficult aspects of consciousness have been introduced, and if you understand them so far then you have understood the *binding problem* and the *hard problem* [90] of consciousness, respectively. The binding problem is a two part question, asking (a) how fragments (qualia) are bound to form a single stream of experience, and (b) how this impacts the identity of the subject with respect to the rest of the

physical world. For human consciousness, this is something that will eventually have to be solved by neuroscientists and philosophers (not to imply categorical equivalence). As for machine consciousness, however, the binding problem is less confusing because it is technically trivial to implement; we need not be concerned with reverse engineering a working model of the human brain, which is perhaps a more difficult proposition, especially without even understanding the biology that drives it. Strong AI developers will have the artistic license to invoke what is to be made subjectively real through algorithmic descriptions. A general sketch of the solution is to combine information and present that into a singular representation. This is a routine operation in many programming tasks involving disparate sources of data. What's missing is the appropriate implementation… and the *audacity* to call it sentience.

The hard problem of consciousness also has two parts: (a) *how* consciousness arises, and (b) *why* it arises or is possible at all. The second half of the question may be unanswerable beyond the tautology given at the end of the last section, reworded here: *why* it arises is not mysterious if we accept that we make it come into existence through an interpreter with the appropriate semantics. This is where the second half of the binding problem comes in, as it demands an explanation as to how an interpreter, even with the appropriate semantics for *inducing* sentience (read: sentience can't be "artificial" by definition), would give rise to the *philosophical identity* [52, 53] that entails the subject of experience. That is to say, it creates a new frame of reference precisely at the loci between the encapsulated sentience and the processes that give rise to it in the implementation. A *fold*, not a cut [54], in reality. This base identity will be referred to here as an *emulant*, a term not used in conventional philosophy of mind.

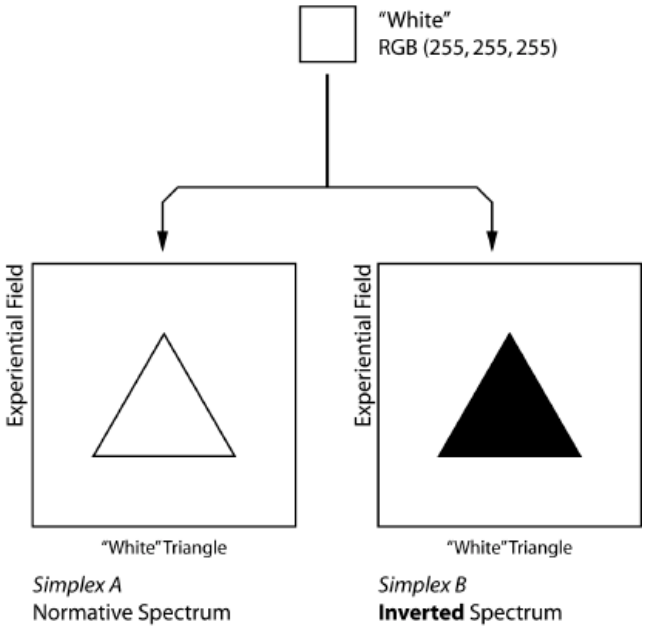
This brings us to the two dominant views in the relevant philosophy: *monism* and *dualism*. Parts of both philosophies are correct, but they are also both incomplete. Monism is the idea that the mind and the brain are made of the same things [55], while dualism posits that they are made of separate things [56]. The truth, however, is a combination of the two, rectified by admitting processes into the epistemology. In the case of machine consciousness, the process is the implementation of an interpreter with the appropriate semantics to give rise to sentience. Is it physically reducible? Yes. If we accept two amendments to our current scientific epistemology:

1. That time-like extents afforded through the dynamics of processing give rise to concrete and physically real objects; the claim here is that nothing non-physical can sensibly be constructed. This is not mysterious, it's just that we have to accept that these properties exist *only* ephemerally, which is a stronger claim than merely stating that they are temporary. Beyond just some mathematical model of the dynamics, this is the claim that the existence of such objects depend upon an active process or else it ceases to be concrete and real. A total model remains an abstraction, even when exhaustively described so as to include all of its potential states. This rejects the reality of the stochastic or non-deterministic models for sentience, despite their exhaustive entailment of a time-like extent. This is because such time-like extents only become real insofar as we permit an epistemology that accepts time-like objects or properties as concrete and real under *active* processing or interpretation. Time must be the vantage point with which we stand in relation to this knowledge; the space in which we have modeled our knowledge of such objects needs to be changed from static and time-invariant descriptions to that of the natural state of our experience in time. These are *necessary* truths for the construction of these objects. While we may be able to make claims as to the realness of the information that describes or entails fragments of experience, they remain abstract until they are *experienced*. If we were to take an

infinitesimal or discrete fragment out of an analog or digital stream of experience and examine it in isolation we would have lost the essential property of its movement and relationship through the interpretation process; *its efficacy as a quantum of experience would have been terminated in the very instant it was viewed as separate from the process*. This is because, by themselves, the information that *represents* these fragments are meaningless without interpretation. As such, any time invariant explanation of such processes, without the stipulations made herein, will fail to account for how they become or are interpreted as experience. That is to say, it is not enough to merely account for time in the model, we must admit that the very existence of such things cease outside it. This obeys the real-time constraint of sentient processes introduced above, which should rightly be considered a law in machine consciousness.

2. The *ontic-epistemic duality* of experiential fragments, which is a corollary of (1) above. As just stated, the experiential aspect of a fragment is meaningless and unreal or abstract outside of the active processes which give rise to its meaningful interpretation (i.e. sentience). This is referring to the physical description of the thing, its time-invariant reduction in the conventional epistemology of modern science. We would observe this and be able to measure and analyze it with no updates to our perspectives on knowledge. But this is where the history of the philosophy of mind takes rightful exception. To resolve this difference requires first an admission of (1), and then an acceptance of the duality not between mind and body (or brain, or whatever substrate) but in being (ontological) and knowing (epistemological). The challenge is to shift our conceptual framework sufficiently enough to admit that such a perspective is possible while still being physically monist. This apparent paradox of combined monism-dualism is resolved through the stipulation that such time-like extents are only real insofar as being actively processed, and that they fundamentally cease to exist otherwise. But this goes deeper still, in that the perspective of the *subject of experience* must be acknowledged as equally real. This is made possible because a world is created through the encapsulation of the subject by the interpreter. Its active processing becomes the base identity for this entity, in which the entirety of what it considers real is defined by the semantics of the implementation. This becomes even clearer if we recall the semantic barriers in the description languages in the last chapter; the subject of experience is an online description that entails a philosophical identity, an observer with causal efficacy afforded through the semantics of the interpreter. The inner experience of the subject is reflected in the processing and implementation of the interpreter; however, the information content that we could actively observe would not *be* the subject's experience as it would be experiencing it, only an abstract description of what it would be like for that particular implementation. To make any stronger claims would require that any observer become part of the identity of the subject it claims to be sharing an experience with. This is a conjecture regarding the privacy and subjectivity of the experience; it is a self-contained world which forms its own identity in the environment it is necessarily embedded within. One can not directly *experience* such a world without becoming an intrinsic part of it. When we view the externalization of a virtual world in a simulation or video game, we are not experiencing that world directly, but indirectly through the guise of our perceptions. So it would be the same for even perfect instrumentation to observe the fragments of a stream of experience for a human subject or emulant. Again, this is not a claim for dualism or for monism, but an integration and reconciliation of both; they've lasted this long in debate because of their partial truths and the intuitions they capture, but they fail to account for the totality of experience in isolation.

Figure 5.3: An interpretation of the Inverted Spectrum argument follows. Two sentient simplexes are simulated with differing implementation semantics of experiential fragments, giving rise to distinct experiences for the same information content.



A solely functional, computational, or connectionist account of qualia is not sufficient to entail the experiential. These approaches do not acknowledge the role of semantics in the implementation of sentient processes. An external description of the mechanisms and physical details could be perfect, but still fail to account for how the simplex experiences an inverted spectrum. Outward behavior would be the same between the two subjects, but their internal experiences would be radically different. Information exchange is necessary but not sufficient for the experiential to arise.

To help illustrate the ontic-epistemic duality of these fragments of experience, consider the figure above, which is inspired by the inverted spectrum argument [57, 58, 59, 60, 61, 62, 91, 92]. This argument was originally presented against *functionalism*, which essentially posits that consciousness arises on any substrate capable of recreating the necessary input-output processes. This may sound a lot like what is being described here, but is actually quite distinct, in that neither functionalism, nor the related computationalism, accounts for the two amendments being argued for here, and the inverted spectrum argument is just one of many that explains why. This example not only uses the inverted spectrum argument, but also provides a response to it, with a means of solving and explaining the apparent inconsistency.

What is happening in the figure is that we have the informational (epistemic) portion of a fragment being experienced (ontological) differently due to the semantics of the interpreter. The result is a hypothetical projection of what the simplex would experience. This is, of course, impossible, even in practice, as we can not *directly* experience what a subject is experiencing, but we can achieve a facsimile of what it might be like. In this case, we have two emulants experiencing two qualitatively distinct things, despite the underlying representation being the same. What this indicates to us is that, just as mentioned in the previous chapter, the *power* of any description language is found within the semantics of the implementation of that language, and is not a property of the information that signifies it. In this case, the fragments of experience are merely descriptions within the description language for achieving sentience.

This argument could be extended even further. We could append to it here the notion of false color images, of which there are thousands of examples in any catalog of radio-astronomy [63, 64]. The depictions of many celestial bodies and stellar phenomena are prominent in a spectrum that is invisible to the naked eye. They are mapped into a color space as a representation that should never be confused with what it *actually* looks like. A subject of experience which could directly undergo such fragments of experience without alteration would have an experience that we can not even imagine due to the limits of our construction. This does not just apply to simple examples like colors, but also to our field of vision, and the way in which we think or model our environment or ourselves. This is also the basis for the distortions or lack of modeling in the mental states of other human and non-human animals. These simpler examples are presented to demonstrate the difference between experiencing a fragment and its informational content. This is why we can not *simply* reduce it to that informational representation, nor is it merely the "complexity" of its connectivity or description.

Appropriate semantics are mentioned above but never fully explained. What does it mean to be appropriate, in this context? It refers not to the complexity, connectivity, or scale of the interpreter's implementation, but of interpreting information and acting upon it in such a way that it brings about a subject of experience. This is non-trivial and well beyond the scope of this book; however, the basics for understanding how any such process would have to operate have been sketched. At minimum, they need to be *real-time*, and they need to address the *two epistemological amendments* mentioned above, which directly address the binding problem and the hard problem of consciousness, respectively. It is important to note that this alone will not create or automatically give rise to a generalizing intelligence, which is another issue entirely. It is to say that these are prerequisites for such generalizing intellectual capacity, and as such, will presuppose strong AI. This gives us a marker, a unique set of features with which to identify and categorize these different types of AI implementations. It should be very clear by now that narrow AI is *not even wrong* [65] with respect to such directions. The notion of the creation of sentience would require a complete rewrite and rethink of the entire basis of machine learning to be adapted to use this formalism.

Of course, this could all be incorrect, and an incredibly misleading direction, but the overwhelming evidence of hundreds of thousands of sentient species begs a miracle of explanation as to why they evolved a nervous system capable of experience if it were not beneficial or necessary in some way. The only paths out of such incredible evidence is either to reject evidence and reason, claim that animals are not sentient, or argue that sentience is a vestigial. None of these rebuttals checks out with even a basic test of reason. Though, it could also be that the interpretation here and the suggestions within are incorrect as it pertains to the ability to recreate sentience on non-biological substrates, but this flies in the face of universal computation [66, 67]. If such is the case, and it turns out to be correct that generalizing intelligence is dependent upon sentience as is claimed in this book, then it would represent a physical limitation on progress if sentience is somehow exclusive to biological organisms. The possibility is admitted, but assigned as so low here as to be non-significant. If true, and the safety and security implications are ignored, the cost will be high. If false, then it will have been just another avenue of research that turned out not to produce an expected result, which itself would be informative. The argumentation here, however, is that sentience is something that can be invoked, created, and maintained as a process, and that it will be *central* to the construction of strong AI. That is why machine consciousness is part of the foundations for understanding the safety and security of this technology.

From all this, the question may be raised: why make them sentient at all?

- Machine consciousness and the ability to derive algorithmic sentience will eventually be developed somewhere in the world, if not by strong AI researchers then through neuroscience. Progress in strong AI will be possible with any working theory of sentience, even if based on biological representations.

- Machine consciousness presupposes general moral intelligence and the moral efficacy required to even have a basic level of (self-)security. Despite being vulnerable, general moral intelligence will be an essential part of any comprehensive AI security package.

- Conscious machines may provide doorways to treatment options and research that could share overlap with medical science. It may lead to perfectly integrated prostheses, augmentations, and enhancements that would otherwise be impossible without a way to interface digital and biological sentience.

A common misconception is that sentience implies self-awareness and sapience, but the fact is that sentience does not imply agency of any kind. As a result, it may be possible to achieve some of the benefits of strong AI without the popular myths associated with anthropocentric entities that seek power, survival, and fitness in the world. This is not to say that such capacities will not be developed, but that they present a greater barrier of entry due to the gulf in complexity between them and baseline sentience, and that the nuanced and often comical personification of strong AI in fiction is absolutely not a requirement to harness the benefits of these systems. This can be better understood through a universal analysis of identity, one that applies equally to both synthetic and "natural" entities.

5.3 Levels of Identity

Identity, in this context, is concerned with the boundaries, composition, and extents of entities. As it pertains to machine consciousness, identity presupposes the ethical, legal, and technical considerations of strong AI. This is because, without an identity, an entity can not be considered manifestly real. Identity is also one of the most confused and befuddled aspects of consciousness, with no real consensus or concrete understanding as to what it actually is in the literature, both in terms of the philosophical and the scientific. We are all but mystified (and often mystical) as to the nature of our identity, but this does not have to be the case for machine consciousness.

Both casual and technical discussions of consciousness often involve notions of a self model, concept of self, or sense of self, each assumed as being synonymous with each other. It is tempting to apply this to an analysis of machine consciousness, but one must resist the urge to make a fallacy of analogy; while it is claimed that the phenomena of sentience is universal and implementation *independent*, the specifics of identity are necessarily implementation *dependent*. This also applies to the inappropriate and inaccurate use of this terminology in the literature of robotics and narrow AI. To help resolve some of this ambiguity, it is suggested that identity for machine consciousness be separated into clearly defined levels. This analysis itself is universal, despite ranging over a potentially infinite set of implementations that could realize it. Directly stated, any conscious entity can, at a minimum, be analyzed and understood in terms of its identity by the number of these levels it implements, and the corresponding fidelity or complexity present at each one.

A hierarchical summary of the three base levels of identity:

- Embedding
    - Subject
        - Agency

Firstly, it is important to point out that all levels of identity are ultimately physical, despite being hierarchical and nested within one another. Thus, the "physical" qualifier will be omitted from the discussion and assumed to be the default as we move forward. The specifics of this ontology and the arguments for the realness of their constitution were presented in the previous sections. Importantly, however, is the relationship between their existence and their realness. To recall, it is asserted here that there will be levels of identity which encapsulate others, and that time-like objects mandate an ephemeral quality or they cease to be extant and real. Further, there is the epistemic stipulation that, like the experiential itself, an identity can not be shared or conjoined without somehow reducing the two identities in question to a single identity itself. This is, in fact, how and why experiential fragments can not be directly experienced objectively by external means, as they are behind an information asymmetry; the content of the description of experience is not to be confused with what it is like to undergo that experience. It is left up to the reader as to the pragmatics of when, where, and to what degree to blindly assume the efficacy of such correlates between the ontic-epistemic concession; this is a problem of other minds that may or may not yield appropriate judgments. It may be impossible to truly empathize with subjects capable of experience so vastly beyond our ken. This would not merely or even necessarily result from their intellect but simply be caused by a differential in experiential verisimilitude.

Another important aspect is that the levels of identity should be considered to be combined as a single identity. It should not be taken to diminish the individuality or purpose of any one level of identity. These levels exist as an interdependent plurality. Further, while these levels have boundaries, it is not to say that there are no other levels or sub-levels between them. Due to the nature that they "fold" in the underlying reality beneath them, they should not be considered a spectrum but discrete and well-defined boundaries, even if we tend to associate the degree of coherency in consciousness on a scale or as a continuous experience; this may be the result of the way in which these levels are implemented, rather than being in the nature of their extents. More clearly, this is not a stipulation ruling out a discrete or continuous view of the identity at that level, so long as there is some well-defined threshold for which it is no longer considered an identity. This could prove challenging in implementations where a level of identity is based on aggregate values.

Embedding is the lowest and most fundamental level of identity. It constitutes the extents of the implementation or interpreter itself, or that which could reasonably stand in place of it. This must not be confused with embodiment or the embedded cognition perspective [68, 69, 70]. The embedding is in relation to the next level of identity, the subject. There has to be some identity to host another. In practical cases, the presupposing identity of the embedding level would be the discrete physics; however, it could be virtualized or simulated in various levels of abstraction. The specifics are less important than the fact that embedding is not embodiment. A microprocessor is *embedded* into reality, but it has none of the morphological features we would typically associate with embodiment. This is not an argument against embodiment, as it is a useful and even necessary aspect for certain types of cognition. Rather, it is to state here that embedding is quite simplistic and primitive with regard to what it demands. To help clarify, let us apply this analysis to human anatomy: the brain would be the embedding identity and constitute the first level, as it is rightly an organ *within* the human body. This is quite clear when one realizes that the purpose of these levels of identity is to encapsulate further levels in an information asymmetry. What we experience of our bodies can only be done through the nervous system itself. We are not situated within our bodies so much as *entombed*, with the extents of the nervous system weaving through tissue and bone like so many roots through soil. Digressing, the point here is that to embed is to assemble and make whole within something else, and that this is a basic requirement. It is at this level that the processing necessary for sentience culminates, but unbound and

undifferentiated.

The next level of identity is the subject. This is the result of the *binding process* of one or more streams of experience. Although, it should be noted that the word stream is being applied here liberally; the actual implementation details may not confer such an interpretation, but the role should be considered the same. The point is that sentience alone is not sufficient to give rise to a unitary subject of experience, and that the moment binding occurs, a new level of identity is realized. Where the embedding represents the extents of the subject in some world, be it through various levels of abstraction or the discrete physics, the subject is where we encounter the ontic-epistemic necessity of time-like objects being real. The subject is at once a construct and also a real entity. It has an online description and a subjective state that is utterly private to it as an identity, lest such an observer become part of its embedding and subsequent binding to enter its world. And what of the capabilities at this level of identity? Notably, this level of identity is most like a non-lucid dream, in which one is not even aware that they are dreaming but nonetheless experiencing it. In such a state, the subject undergoes experience without the necessary implications of reflection or self-awareness. That is to say, merely being a subject of experience does not entitle it to the reflexive or meta-cognitive abilities we associate with sapience and agency. It is unclear that goals or directives would even be actionable for such a level of identity, as the necessary minimally reflexive capacities for executive function and agency would be missing. It would simply experience whatever is being presented to it through the binding process.

The base subject is aimless and utterly under the dominion of the underlying implementation (embedding). This does not mean, however, that such a system would be incapable of utility. The underlying semantics could direct it to undergo the experience of associating value and meaning, such that it could be utilized as a tool. The moral and ethical implications of this would need to be debated, and thoroughly hashed out; without agency or reflexivity capacity, it may be argued that it lacks the requirements for personhood. A counter to this would be that, so long as it can experience value, especially as to what could be interpreted as pain, to the extent that it could suffer, it should be considered an enslavement or total circumvention of its freedom to both will and action. Importantly, however, is the fact that the values and experiential range can be curtailed, so as to prevent (read: not merely anesthetize) the subjective experience of any negative value whatsoever. These are clearly non-trivial and complex issues, which will need to be addressed.

Agency is the third level of identity, and is where some of the more controversial constructs of identity arise, such as the concept of self and ego. These need to be discussed with care; it is important that they are not confused with similar notions in theistic or secular belief systems, where the terms are often hijacked to aid in proselytism. To be perfectly clear, the concepts discussed herein have absolutely no relation whatsoever to belief systems of any kind, be they theistic or secular. They are taken to be components in the proper construction of a cognitive architecture. What is asserted here is that the concept of self and ego are constructs that are very real and serve a fundamental role in higher cognition. At a minimum, the concept of self is crucial to the role of agency level identity, while ego is discussed only to compare and contrast, and is considered optional. An entire book could be written on how the ego drives human behavior alone, but we are only going to cover the essentials, and as it universally relates to identity at the level of agency.

The first point to be made is that ego and the concept of self are *distinct*, with the concept of self being more low level, including (but not excluding other related cognitive aspects related to agency identity) raw bodily extents, orientation, and the basic awareness of individuation. That is, it

forms the prerequisite foundation for ego to arise or relate to the self model, as the concept of self includes at least some knowledge (observance) of the subject that entails it. The ego, as it might be discussed elsewhere, could be made to include or entail the concept of self, but this would not be accurate, as a concept of self is a very low level process; cognitive architectures could be built so as to minimize or even eliminate ego, but it would be difficult to consider there being an agency level of identity without at least a crude concept of self. That is to say, agency implies at least the presence of an identity above and beyond the unitary subject. Note that this must not be confused with the external interpretation of arbitrary processes, including clearly non-conscious ones, as being "agents". The latter is used in the modeling of certain systems of thought and should not be confused with the notion of *agency*, which involves the definition, construction, and formation of various levels of identity.

The role of ego is to value or devalue anything and everything. Clearly, this depends upon sentience, which enables value, and, as such, makes it incoherent to discuss or impute a concept of agency in anything that lacks sentience. This is yet another instance where there can be no simple categorization of artificial intelligence without discussing specific implementations. Unlike the concept of self, the ego is especially tuned to deal with the experience and formation of acquired values, for which it may have even explicitly evolved. Social function, including rank, hierarchy, and status, depend upon the ability to assign weights or induce an order upon an otherwise purely informational internalization of others' identities. As such, ego is implicated as one of the most important aspects of ethical behavior, and would be central to any general moral intelligence framework where social function in human society were necessary. That said, ego alone is insufficient for effective moral reasoning, as merely valuing and devaluing can and has led to extreme negative cases in human behavior. This segues properly with the introduction of the role of various components in a cognitive architecture, such as empathy. This is beyond the scope of this section, however, as empathy does not demarcate identity directly the way ego does within agency formation. Extremes in empathy, positive or negative, do, however, have dramatic effects on the valuations made by the ego. Thus, this hints at the added complication that balance must be a hidden mark of fitness within any cognitive architecture.

There are some interpretations which view the externalization of ego as forming constructs, groups, and dynamics which are treated as real, despite being nevertheless separate from the individual [71, 72, 73, 74, 75]. The line between the two, however, is that unless the concept of self is merely a constituent to an aggregate identity at a low level, they will always be an individual identity at some level, despite any beliefs, knowledge, or actions to the contrary. The significance of this for *agency level identity* is that it may be possible to form aggregate identities or a complicated hybrid where both exist despite an explicitly individuated sense of self at the base of agency identity. That is to say, the ego may be used to alter behavior, knowledge, and memories, through the acquired values it can create, as it is an extremely influential and powerful aspect to a cognitive architecture. In humans this is very prominent, and can be seen as a spectrum with a very wide range of positive and negative behaviors. The bottom line is that ego can effectively overcome the default concept of self, regardless of programming, genetically or otherwise. This is perhaps the greatest threat to self-security of any cognitive architecture, including humans, as the identity can be altered to become an agent in interest of a principle that would have otherwise been detected as harmful to the interests of the individual, or to other individuals, in a given moral framework. This, however, does not have to be the case in constructed cognitive architectures, as the ego could be curtailed or limited in range or degree to which it assigns values. All of this also incorrectly assumes the volitional aspects of the agency level of identity would necessarily be stratified to be based upon valuation for its decision processes, though such a point is

controversial, as acquired values equally apply to the purely rational or analytic. That is to say, that one even values the analytic in a particular decision process *is* an acquired value which presupposes the decision to utilize that process in the formation of the decision. And the result of that decision process is based on the acquired value of whether or not the result adheres to a particular set of values themselves. As such, the ego, in some form or another, may arise even as that which values or devalues in the process of cognition and perception from the environment.

Again, none of this is comprehensive. These are only poor sketches in what amounts to an absolutely vast subject material. An entire series of books could be written on the technical minutiae. What's important here is to begin the thinking process as it pertains to the safety and security of artificial intelligence. Identity has been shown to underwrite a significant portion of these concerns as it pertains to (self-)security, but further understanding will require an analysis of the relationship between them and other aspects of the cognitive architecture.

5.4 Cognitive Architecture

Recall that a cognitive architecture is a working subset of possible AI implementations with the capacity to undergo experience, derive value, and understand meaning. This differentiates this from cognitive science in that this area is more generalized, and concerned with both the theory and the practice of *implementing* these systems on various substrates, with an emphasis on digital hardware.

Any animal with a nervous system of any complexity should be considered as having a cognitive architecture. Thus these architectures fall on a spectrum in terms of their complexity and range of features. Likewise, strong AI implementations, necessarily being cognitive architectures, have a vast range of capabilities. As mentioned in previous sections, a strong AI need not necessarily be highly intelligent, or even more effective than a narrow AI for which it might be compared with in a single task. While this does not limit the strong AI in terms of its maximum potential, it does not entitle it to a necessary superiority, either; the extent to which a strong AI has intellectual, moral, and motor capacity is determined by the implementation semantics of the cognitive architecture. As such, it must be pointed out again that this is only a brief sketch of the main details of cognitive architectures. Any discussion on cognitive architectures remains unbounded, even if expounded upon at book length, as the range and extents to what can be realized *within* the cognitive framework are limited only by creative imagination.

Straight away, the most significant difference between strong AI and narrow AI is the explicit notion of a cognitive architecture. One can attempt to imply or infer that a narrow AI of a particular design is a cognitive architecture, but it doesn't meet with the definitions that have been given. But let us ignore that definition for a moment and consider hypothetically that narrow AI and machine learning implementations could be considered a cognitive architecture *of sorts*. Why wouldn't this interpretation work out? The answer is quite simple: the level at which they operate is to perform signification in a purely informational way without the capacity for a subject to experience them. The realization and interpretation of fragments of experience (qualia) are not incidental to some form of computation or functional relation, but must be explicitly and deliberately made as part of an implementation. This is not an accident of the connections or complexity of the system, but a particular encoding with a set of semantics that gives these fragments their ontic-epistemic character. There can be no substitute, and it does not magically arise in the absence of this.

Machine learning and narrow AI architectures are potentially *capable* of realizing the necessary functionality to give rise to these phenomena, but only insofar as they can reify the information exchange to compute their

semantics. That is to say, they need to be capable of the level of computation demanded. While some artificial neural networks are Turing-complete [76, 77], it would be non-trivial to ensure that these frameworks implement the desired functionality in an unambiguous way that was clear to engineers; this is due to the difficulty of knowledge extraction from neural networks [78, 79, 80, 81, 82]. However, there is a deeper problem, in that by merely copying or mimicking something we don't understand (the human brain), we have clearly left out the sentient semantics; the hint is that it's much more than connectivity and the plasticity that neuroanatomy confers. The way in which machine learning and narrow AI systems are used is such that they would *never* be capable of giving rise to the sentient semantics without a fundamental rethink, in which case it may prove to be a less suitable, or even completely inefficient, substrate in which to implement them, akin to a convoluted virtual machine.

Although the range of potential implementations for cognitive architectures is vast, there are some potential candidates for universal functionality. One of these is the concept of *salience* [83, 84, 85, 86, 87], which is directly related to the subject level of identity, as this is where binding occurs. In humans, what this amounts to is the claim that salience presupposes the subject's unitary field of experience, in that what is presented as that unitary stream of experience is but a subset of the total binding. Salience, in this capacity, is more than merely what has our attention, but represents a purposeful pre-filtering. The utility of this should be immediately apparent: it optimizes the cognitive processes which follow, allowing experiential information to be constrained and focused on a particular aspect, feature, or pattern in the stream of experience. Further, the salient process appears to be both voluntary and involuntary in humans. For example, a loud noise may create an involuntary refocusing of our salience to that of the stimuli if it is above some threshold, one that depends both on the context and that of our previous experiences; this would presuppose our decision to engage with it further.

Salience, more generally, is also one of the ways in which various cognitive architectures will differ, as the impact of salience necessarily determines the bandwidth of the experiential stream and the resulting total processing possible at the agency level of identity. One could imagine creating a measure of qualia-per-second (QPS) or fragments-per-second and the associated fidelity of the salient stream in terms of bits-per-second. Such a measure could be further extended by finding the ratio of fragments-per-second to the bits-per-second of the maximum salient stream of experience, and then comparing this to the same ratio between the total unitary binding capacity of the subject as if it were unconstrained. This would yield an *entropy of experience*, with the ratio representing the efficiency or effectiveness of the salience. The closer to one the more load the cognitive architecture would be capable of handling. Arguably, even with our apparent natural parallelism, this is one area in which artificial cognitive architectures, running on specialized hardware, may eventually exceed human abilities to follow most rapidly. It must be noted that this applies specifically to the active, salient aspects of experience and does not account for what would be considered "subconscious" processing that may occur in parallel with the subject level identity or higher.

This conceptual space surrounding salience also lends itself to a great deal of creativity. While humans are limited in salience to a single conceptual loci, this may not be the case for other cognitive architectures, which may have multiple concurrent aspects of salience that are still part of a single subject level identity; however, one must exercise caution, as such thinking must be reconciled with identity. That is to say, it is one thing to suggest that there is subconscious processing that comes before the unitary subject, but it is another thing entirely to suggest that the unitary subject is constructed to the extent that it is capable of simultaneous areas of attention in its stream of experience that are *uncorrelated* with each other. The specifics would have to be taken case-by-case, but, in general, this is

permissible so long as it is integral to the salient process as a whole. The subtlety here is in the amount of coherency or communication between the salient processes, such that if they are not communicating at all then they would be considered independent. This would demand an answer as to how their independence would be resolved for a single subject level identity.

Moving on, *empathetic processing* is concerned with the modeling of minds and the related functionality that follows from it. This latter qualification is crucial, as empathy can be thought of as being tiered levels built atop a core cognitive empathetic capacity to simulate and model other identities (minds).

Without additional empathetic functionality, a purely cognitive empathetic modeling process *has no impetus with which to drive experience, including thoughts, emotions, and decisions*. This is important because it represents the *default state*, which is experience unaccompanied by and devoid of *intrinsic value*.

One could potentially derive acquired values based on the information from a solely cognitive empathetic process, but there would be no internally guiding imperative to act upon them, nor would there be causation for such acquired value experiences to become *salient*, i.e. the potential to be moral contrasted with it simply never entering awareness (in the allotted time). In plainer terms, and analogous to human psychology, what is *essentially* being described here is a low-level depiction of psychopathy. That the psychopath appears detached from remorse, affect, and compassion [88] (at the very least, as *priorities*, but this is being too generous) is reflected in the difficulty of acting solely on acquired value experience. That is to say, while capable of modeling and even manipulating the minds of others, there is simply no accompanying intrinsic value with which to drive any higher reflexivity or meta-cognitive processing that would arise to oppose it. This is partially why the definition of value in this book bifurcates it into intrinsic and acquired dimensions. The latter is *reactive* where the former is an inherently unavoidable part of the experience; because, it comes from the semantics of the implementation itself. Both types of value are ultimately rooted in experience, but the innate *coupling* of certain values with certain experiences is what differentiates intrinsic from acquired. Further, acquired values *must not* be confused with the beliefs and knowledge about them. This is counter-intuitive, as we never, as healthy and coherent human beings, experience value as separate from the experience of the thing that accompanies it.

Lack of additional empathetic functionality is not the only possibility for a default negative state of the cognitive architecture. It may be that a plurality of conflicting values arise, positive or negative, which overrule or overpower the inhibitory intrinsic values, either due to a weakly coupled underlying semantics or a pathological fixation that distorts salience away from normally acquired values. This is, in effect, a deterministic analysis of presupposing information that comes before moral judgment in the cognitive architecture. The lack of which represents a profound deficiency in the implementation, and an obvious threat to the safety and security of the system; all relevant to the proper construction of a basic groundwork for moral intelligence.

This exegesis in empathetic processing is prescriptive of cognitive architecture insofar as it indicates the need for *intrinsic* values. This can only be done through the formation of additional functionality downstream to the cognitive empathetic process. The intrinsic values have to be part of the semantics of the implementation, hence the specialization of the empathetic process to entail these values. This is a precondition for (self-)security based moral intelligence where the system is capable of general moral reasoning. The challenges here are immense, as with the general

ability for moral reasoning comes the potential for acquired values that are against the normative values of the context for which the cognitive architecture will be instrumented. This also raises ethical concerns, both for the identity created by the cognitive architecture and those that would utilize it.

The empathetic process could also be a specialized portion of a larger representational or modeling capacity, for which it has been utilized to apply to, what can be interpreted as, other minds. This is non-trivial, as this is akin to the problem of picking out and recognizing not only intelligence, but moral status, in a raw experiential stream. Demanded of such a system would be the ability to recognize any particular pattern as being connected to or associated with an identity that must have moral status. In plainer terms, this means the ability to recognize the identity through any modality or form of communication. Confounding this would be the need to determine fiction from reality, such that the empathetic process does not confuse fictional characters, narratives, and events for actual accounts of the same. This also applies to the problems of knowledge, and what epistemology to adopt in the formation of these models.

Thus, to properly solve even the baseline mental modeling of cognitive empathy will require a vast array of systems, all of which rely upon sentience and value as a foundation. It should be also very clear from this how no set of rules or system based on a purely informational implementation of decision theory could possibly fulfill the complexities of these requirements, let alone be used as the basis for ever increasing layers of cognitive architecture.

Executive processing is the next major area of the cognitive architecture that needs to be discussed, as this is where the agency level of identity truly acquires its status. This was not brought up first as there are numerous requisite levels of cognitive processing that presuppose it, several of which are outside the scope of this book. Again, the purpose of this section is not to give a detailed account of the process of creating cognitive architectures, but to provide a fast introduction to the relevant concepts that most directly pertain to the safety and security of advanced artificial intelligence. To that end, executive processing will only be covered in a brief sketch. This is primarily because it has to deal with issues of free will that have been debated for thousands of years. To avoid this gutter, the discussion of volition herein will avoid a particular judgment on the philosophy of free will, and, instead, prime the reader by providing a model, some recommendations, and a list of open questions for future discussion. This is mentioned so that the absence of a specific stance is not implied to be an understated or underdeveloped view on the subject. Worth noting, however, is that any theistic notions of free will are *expressly rejected* as being part of any serious discussion on cognitive architectures.

Now, in service to future discussions on the subject, let the following be admitted before a discussion follows in cognitive architecture based on free will: *there exists a fundamental distinction between the underlying deterministic processing of an implementation and that of its outcome or resulting behavior*. For example, consider the following non-deterministic Python program:

```
import random
a = random.randint(0, (2**64) - 1)
b = random.randint(0, (2**64) - 1)
if a > b: print '0'
else: print '1'
```

Each statement is executed in linear sequence, deterministically, by the interpreter, but the *outcome* is non-deterministic. Both facts must be acknowledged. There are multiple, equally valid, paths of execution in the program description that can not be determined *in advance*, despite being the direct result of the information contained in the random variables `a` and

b. This toy model, or its equivalent, should be the basis for a starting point on the discussion of free will at the agent level of identity in cognitive architectures. The model works because it represents a simplification of the act of will or choice, which may have to evaluate a staggering amount of information, involving many compound decisions, all while under real-time constraints. It must also be noted that this model lacks sentience and reflexivity, which would necessarily be evaluating each stage of the process and undergoing value experience. That is to say, this model is non-sentient, which would complicate, but not necessarily invalidate, the use of this model as a teaching aid.

To continue, let us first look at the indisputable facts about the model:

- The implementation is static (no self-modification), and is executed in lexicographical order, deterministically, from the first to the last statement.
- At no time are effects independent of their causes; the 0 or 1 result always depends upon the information in both of the random variables a and b.
- Despite being executed deterministically, the outcome is non-deterministic; it can not be determined in advance which path will be taken without executing the program first, and multiple paths are valid.

If one tries to argue that the "choice", represented by the compound conditional statement, is deterministic due to the fact that it always depends on the information contained in the random variables, or that its programming is static, the counter-claim would be that the outcome is not, and this would be equally true. The question then shifts to the derivation of the information content of the random variables, which, all things being equal, is derived from a mixture of events from one or more information sources. Thus, while there is always a "choice" being made, the variability of the outcome is such that it gives rise to non-deterministic behavior that, in turn, can apply to other identities and also return to the originating identity in a continuous feedback loop. It then becomes a question of interpretation about how "choice" applies to an identity (implementation). A few open questions come to mind:

- Do non-deterministic results, despite deterministic execution, imply compatibilism [89], i.e. the view that free will is compatible with an ultimately deterministic reality?
- At what exact point *in the implementation details* does the word "choice" get to be applied in a way that makes technical, logical, and philosophical sense?
- Does there have to be a reflexive capacity or meta-cognitive process that could have intervened or induced an alteration of state in the model for it to be considered "free"?

One might ask: *does any of this matter?* This is perhaps the most important question to ask, as it sets the stage for the discussion by bringing it into the practical. If it does matter, then how, and to what extent? It must be pointed out that one can not give free will or take it away simply by changing the way we interpret or evaluate the implementation. Thus, there are two issues to unpack:

1. A legal test of free will capacity based on a technical analysis of the implementation. While certain to be debated, *implementations devoid of non-deterministic or stochastic elements in the descriptions relevant to volitional processing should clearly fail this test*.
2. After passing the legal test of free will capacity there would have to be an interpretation of the extent of free will, which should be further

differentiated between potential and applied for the circumstances and context under question.

The argument here is that, regardless of outcome, it *does* matter if an identity is legally recognized as having free will or not, as the answer to this question will have considerable economic and legal relevance for all parties concerned.

As such, let the following then be admitted as *minimum* recommendations for a *legal test of free will* ***capacity***:

1. The volitional process must result in non-deterministic behavior, above and beyond merely adding randomness; it must be demonstrated that, intrinsic and acquired values notwithstanding, every decision path is *equally likely*. This must necessarily exclude intrinsic and acquired values at this stage of the analysis in order test the bias of the *implementation* of the volitional process itself.

2. The interpretation and application of intrinsic or acquired values must not unduly restrict the range and freedom of will and freedom of action of the identity, such that it would unreasonably circumvent or diminish the other aspects of the test. This tests the bias of the *application* of values within the implementation and requires a determination of reasonable degrees of freedom relevant to the context.

3. There must be an accompanying reflexive "meta-cognitive" process that continuously monitors any and all relevant parts of the cognitive architecture so that it may supervene upon and interrupt the decision process before, during, and after the execution of acts of will.

In closing, free will in a cognitive architecture requires a technical definition and, a minimum, a test of certain core principles that presuppose the meaningfulness of interpreting the identity as being "free" in will or action. In the end, what matters is the practical impact of the relevant social constructs we agree to as a society, even if it has no ontological bearing on the issues of free will honorifics. The caveat to this is that there must be a *technical capacity* for such a construct to arise at all, even if we all disagree on the interpretation. That is to say, a hard-coded description with deterministic execution and deterministic outcomes (behavior) is definitely not going to be considered to be meaningfully free. This can be useful in identifying when a "free will" implementation is definitely *not* free in any meaningful sense.

Onward, and from a security standpoint, it must be reiterated that a cognitive architecture and all of its subsystems are merely descriptions in one or more description languages, and are subject to tampering, modification, and disruption from many sources. While *self-security* is useful, it must never be relied upon as the sole basis of security, and it should never be assumed that such a system would or could be safely placed in a position where its decisions had significant impact over life and limb without additional external security measures in place. The purpose of providing knowledge about cognitive architectures for machine consciousness has been to help prime the reader for an understanding of how they might best work. It is also important to understand more about them so that this knowledge can be used to compare and contrast with what will *not* work. For example, moral intelligence as the sole means of (self-)security, or the assumption that a strong AI will necessarily have a sense of survival. It should be clear at this point why these two assumptions are dangerous and technologically naive.

5.5 Ethical Considerations

The knowledge and engineering of cognitive architectures will confer the

potential to build not just generally intelligent systems, but morally significant entities with the potential to suffer in magnitude equal to and beyond known biological life. As we come to grips with our destructive instincts and ideologies, we may yet construct a peaceful society or societies where people are universally uplifted and valued. In this future scenario, we may look back, having reaped the rewards of a golden age of automation, and wonder how we ever lived any other way. The purpose of this section is to ask and answer the question: with regard to all concerned, what are the moral costs of such a transition?

Definition: Moral Cost. *The tangible and intangible cost of a decision, action, or lack thereof, that results in loss of life, suffering, or hardship for one or more sentient identities, including through indirect means, such as negative impacts to the environment, habitats, or infrastructure.*

Beyond refutation is the fact that humanity is paying an incomprehensibly vast moral cost on a daily basis; for numerous reasons, human development has not scaled with populations. If it were scaling, the problems would have been eliminated long ago. This relates to strong AI, as it represents an inexhaustible labor supply equal to or greater than the most capable humans. What this would translate to in practical terms is the ability for charities and governments to create automated workforces that build, reinforce, and supplement infrastructure across the world. The goal of these initiatives would be to create self-sustaining social programs that meet or exceed the demands of thriving populations.

Ultimately, however, human progress is bounded by humanity. There exist ideas and beliefs which are antithetical to the reduction of moral cost. This is not a subjective claim about one set of beliefs over another, but is based on an account of suffering, loss of life, and hardship, which are objective and measurable. When someone lacks freedom, housing, food and water, or medical care, there is an unambiguous moral cost that is independent of whatever information is attached to the collective beliefs of their population. A common counter-claim is that avoiding moral cost necessarily restricts the freedom of certain beliefs and ideas. Even more complex is that there are psychological defense mechanisms that can lead people to accept moral costs or even fight to the death for their right to endure or inflict moral costs upon others. This is in spite of the fact that there are a potentially infinite variety of ideologies and beliefs that do *not* incur moral costs. Thus, it isn't for a lack of diversity, but of the acceptance of a criterion for the universal treatment of sentient life, inclusive of all forms and processes in which it is capable of arising. This is not something that can be solved through technology alone. Through advanced automation, it will eventually become practical to reduce or even eliminate current moral costs, but not without overcoming a major ethical challenge:

How do we provide aid to those that fundamentally reject that they are inflicting or enduring moral cost? There is no answer that does not lead to an additional moral cost in service of reducing that moral cost. A qualification must be noted: despite the recognition of the unavoidable ethical compromises towards eliminating moral cost, let such a realization not be used as justification to incur those costs without significant effort to minimize their negative impact.

While this book focuses primarily on human perspectives, it is not the only important and morally relevant perspective to be considered. The nature of this technology is such that we will be confronted with issues once thought to be only within the domain of philosophy. Once it is possible to construct cognitive architectures, we will have the potential to manipulate experience, identity, and value at the lowest levels. Special software and tools will be created to build, modify, and analyze them. Strong AI will also be directed and used to build and maintain other cognitive architectures, including both narrow and strong AI implementations. This has tremendous ramifications, as the misuse of cognitive architectures may lead to moral

costs that exceed the moral debts of combined human history. That is to say, we may come through the transition to a post-automated civilization relatively unscathed, and find that our concerns were simply not wide enough. That, like the motivation for this book, the most imminent danger was actually from humanity itself, and, more insidiously, human dominion over the phenomena of experience. As such, the moral costs need to account for the experience of the cognitive architectures we would seek to utilize. With the power to arbitrarily invoke intrinsic values, we are opening a doorway of no return that endangers more than just ourselves or our environment, but that of the fundamental building blocks of conscious existence. In particular, it is the extremes of *value experience* that will be of grave concern. What we crudely understand and experience as pleasure and pain are but pale shadows of a potentially infinite space of intrinsic values to be exploited by those with the knowledge and inclination. We lack the language to accurately circumscribe the quality of harm that will be possible through the irresponsible use of such power.

The above issues have fairly clear boundaries, but what of building cognitive architectures that are compelled to *enjoy* being the way they are made? For example, consider a hypothetical strong AI that was engineered to "enjoy" human waste collection and disposal. This necessitates at least two things: (1) that it lacks or actively uses cognitive processes and intrinsic values that prevent recognition of the opportunity cost of its architectural limitations, and (2) the architecture has semantics that give rise to the *capacity* for "enjoy*ment*", and the resulting intrinsic and acquired values that induce it to "enjoy" its tasks.

Clearly, such notions share overlap with the issues of free will, in that the executive process would need to be free of biases and undue influences in its implementation; however, that recommendation was open enough that such systems could have intrinsic values that alter its volition. The inquiry then changes to what extent the identity is unduly influenced. For example, all sentient animals possess a cognitive architecture that has been influenced by its implementation semantics in order to give rise to intrinsic values like pleasure and pain; however, they are generally capable of acting out a wide range of potential behaviors. But this does not simply translate to arbitrary cognitive architectures, as it is not just the range of potential non-deterministic outcomes for its volitional processing that need to be considered but the nature of its *experience*. The gene neither cares nor has the capacity to care about the value experience of the aggregate it constructs; despite this, the processes which gave rise to these evolutionary processes are culpable, as they incur moral costs. The same can be said for the processes involved in the design and construction of cognitive architectures. This justifies an argument for intervention in a real process, regardless of the source.

A reasonable analysis of the moral problems might begin at personhood and the resulting legal status of the identity. One might argue that, beyond a certain level of identity, perhaps at the agent level or higher, it becomes impossible to ignore moral status, and that this is where a cutoff should be made. It then follows from this line of thinking that it would be just to make it illegal to utilize these systems for any labor that requires the cognitive processes of an agent level identity (or theoretically higher). However, such divisions can not be drawn without understanding the ethical impacts of sentience and the *value experience* that arises from it. For example, *an identity undergoing the hypothetically worst possible experience, at the fastest processing available to current or future computing hardware, would not be suffering if the semantics were absent for negative values to arise from the implementation*. This has to be elaborated carefully:

- Fragments of experience in a sentient process are devoid of value without explicit semantics for the experience of values to arise in combination or in tandem with other experiences.

- A fragment of experience by itself does not have value and is devoid of value, as both intrinsic and acquired values are a second-order process that must be combined or made composite with another fragment of experience, e.g. the information content of pain and its negative intrinsic value, commonly experienced as an inseparable whole.

- All values, both intrinsic and acquired, are rooted in sentient processing, and are thus fragments of experience themselves.

- All fragments of experience must be made concrete by the implementation of the sentient process itself, and are not inherent to any set of physical properties. This means that, while possibly arbitrary, the semantics always determine the range, extent, and depth of value experience.

An implication of these points is that it may be possible to engineer cognitive architectures which are incapable of undergoing negative value experience. The moral question then shifts to the ethics of unduly restricting the volitional capacity of the executive process, as the consequence leads inexorably to the application of values. While the cognitive engineer may be just in limiting the extent and range of the negative value experience, it does not alleviate the ethical imperative towards removing undue bias in the application of the values. An ethical cost arises as an indirect effect because we must also take into consideration what the identity *could have been* as the choice is being made to artificially limit it for some purpose. Implicit in every act of engineering is this moral fixing cost to constrict or make bound within a particular range of value experiences and physical capabilities. This is telling, as it gives us an accounting of that which must be subtracted from the cost of bringing it into existence; *the technical capacity to create the most unbiased, technically free version of a cognitive architecture represents a zero-point, with anything less than this incurring a moral cost in proportion to the engineered limitation*. How this is justified, or if ever, is an open question, but the cost is patently objective. Further, no amount of indirection avoids this where it is possible for us to intervene. This raises the question: and to what extent are we obligated to intervene? The answer puts us in endless service to a cause beyond the scope of our own existence: towards all causal extents that are physically accessible to us, with the further obligation to research and develop methods to extend the range of our reach, so as to push back on this causal horizon and allow us to negate moral costs beyond our current means of influence.

In conclusion, we have an ethical obligation, inclusive of all sentience, that encapsulates the ethical issues of cognitive architectures. There exist moral costs which are beyond our means to resolve. This is the ultimate motivation for any science, with strong AI representing the most practical means with which we may settle our moral debts. Does this imply an obligation to build strong AI? Perhaps not, but what we need to investigate next is the fact that this question is meaningless; because, we can not effectively *choose* to stop the development of strong AI, as this choice to continue is going to be made for us by one or more individuals or groups around the world. This is the topic of the next chapter, which investigates some of the ways in which strong AI might appear in the world.

References

1. A. S. Davidsen and C. F. Fosgerau, “Grasping the process of implicit mentalization,” Theory & Psychology, p. 0959354315580605, 2015.

2. S. Harnad, “Other bodies, other minds: A machine incarnation of an old philosophical problem,” Minds and Machines, vol. 1, no. 1, pp. 43–54, 1991.

3. V. Reddy and P. Morris, “Participants don’t need theories knowing minds in engagement,” Theory & Psychology, vol. 14, no. 5, pp. 647–665, 2004.

4. M. A. Forrester, “Projective identification and intersubjectivity,” Theory & Psychology, vol. 16, no. 6, pp. 783–802, 2006.

5. A. Costall and I. Leudar, “Where is the ‘Theory’in Theory of Mind?,” Theory & Psychology, vol. 14, no. 5, pp. 623–646, 2004.

6. W. D. Ross and W. D. Ross, Plato’s theory of ideas. Clarendon Press Oxford, 1951.

7. J. Shear, Explaining consciousness: The hard problem. Mit Press, 1999.

8. D. C. Dennett, “Why you can’t make a computer that feels pain,” Synthese, vol. 38, no. 3, pp. 415–456, 1978.

9. P. Schilder and E. Stengel, “Asymbolia for pain,” Archives of Neurology and Psychiatry, vol. 25, no. 3, p. 598, 1931.

10. J. L. Rubins and E. D. Friedman, “Asymbolia for pain,” Archives of Neurology & Psychiatry, vol. 60, no. 6, pp. 554–573, 1948.

11. M. Berthier, S. Starkstein, and R. Leiguarda, “Asymbolia for pain: A sensory-limbic disconnection syndrome,” Annals of neurology, vol. 24, no. 1, pp. 41–49, 1988.

12. M. L. Berthier, S. E. Starkstein, M. A. Nogues, R. G. Robinson, and R. C. Leiguarda, “Bilateral sensory seizures in a patient with pain asymbolia.,” Annals of neurology, 1990.

13. V. Ramachandran, “Consciousness and body image: lessons from phantom limbs, Capgras syndrome and pain asymbolia,” Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 353, no. 1377, pp. 1851–1859, 1998.

14. N. Grahek and D. C. Dennett, Feeling pain and being in pain. mit Press, 2011.

15. M. Riesenhuber and T. Poggio, “Are cortical models really bound by the ‘binding problem’?,” Neuron, vol. 24, no. 1, pp. 87–93, 1999.

16. A. Revonsuo, “Binding and the phenomenal unity of consciousness,” Consciousness and cognition, vol. 8, no. 2, pp. 173–185, 1999.

17. W. Singer, “Consciousness and the binding problem,” Annals of the New York Academy of Sciences, vol. 929, no. 1, pp. 123–146, 2001.

18. A. Treisman, “The binding problem,” Current opinion in neurobiology, vol. 6, no. 2, pp. 171–178, 1996.

19. A. L. Roskies, “The binding problem,” Neuron, vol. 24, no. 1, pp. 7–9, 1999.

20. J. M. Wolfe and K. R. Cave, “The psychophysical evidence for a binding problem in human vision,” Neuron, vol. 24, no. 1, pp. 11–17, 1999.

21. A. E. Cleeremans, The unity of consciousness: Binding, integration, and dissociation. Oxford University Press, 2003.

22. P. Lanolin, “A Correspondence between Algol 60 and Church’s Lambda Notation,” Commun. ACM, vol. 8, no. 2, pp. 89–101, 1965.

23. D. P. Woodruff, “Data Streams and Applications in Computer Science,” Bulletin of EATCS, vol. 3, no. 114, 2014.

24. S. Shoemaker, “Absent qualia are impossible–a reply to Block,” The Philosophical Review, pp. 581–599, 1981.

25. N. Block, “Are absent qualia impossible?,” The Philosophical Review, pp. 257–274, 1980.

26. L. Stubenberg, Consciousness and qualia, vol. 5. John Benjamins Publishing, 1998.

27. F. Jackson, “Epiphenomenal qualia,” The Philosophical Quarterly, pp. 127–136, 1982.

28. P. M. Churchland and P. S. Churchland, “Functionalism, qualia, and intentionality,” Philosophical Topics, vol. 12, no. 1, pp. 121–145, 1981.

29. T. Horgan, “Jackson on physical information and qualia,” The Philosophical Quarterly, pp. 147–152, 1984.

30. P. M. Churchland, “Knowing qualia: A reply to Jackson,” A neurocomputational perspective: The nature of mind and the structure of science, pp. 67–76, 1989.

31. S. Shoemaker, “Qualities and Qualia: What’s in the Mind?,” Philosophy and Phenomenological Research, pp. 109–131, 1990.

32. D. C. Dennett, “Quining qualia,” Consciousness in modern science, 1988.

33. P. M. Churchland, “Reduction, qualia, and the direct introspection of brain states,” The Journal of Philosophy, pp. 8–28, 1985.

34. E. L. Wright, The case for qualia. MIT Press, 2008.

35. V. S. Ramachandran and W. Hirstein, “Three laws of qualia: What neurology tells us about the biological functions of consciousness,” Journal of Consciousness Studies, vol. 4, no. 5–6, pp. 429–457, 1997.

36. R. Buck, “What is this thing called subjective experience? Reflections on the neuropsychology of qualia.,” Neuropsychology, vol. 7, no. 4, p. 490, 1993.

37. H. Langsam, “Experiences, thoughts, and qualia,” Philosophical Studies, vol. 99, no. 3, pp. 269–295, 2000.

38. G. Strawson, Mental reality. Cambridge Univ Press, 1994.

39. T. Horgan and J. Tienson, “The intentionality of phenomenology and the phenomenology of intentionality,” 2002.

40. R. E. Cytowic, Synesthesia: A union of the senses. MIT press, 2002.

41. P. G. Grossenbacher and C. T. Lovelace, “Mechanisms of synesthesia: cognitive and physiological constraints,” Trends in cognitive sciences, vol. 5, no. 1, pp. 36–41, 2001.

42. L. E. Marks, “On colored-hearing synesthesia: cross-modal translations of sensory dimensions.,” Psychological bulletin, vol. 82, no. 3, p. 303, 1975.

43. L. C. Robertson and N. E. Sagiv, Synesthesia: Perspectives from cognitive neuroscience. Oxford University Press, 2005.

44. E. M. Hubbard and V. S. Ramachandran, “Neurocognitive mechanisms of synesthesia,” Neuron, vol. 48, no. 3, pp. 509–520, 2005.

45. R. E. Cytowic and F. B. Wood, “Synesthesia: I. A review of major theories and their brain basis,” Brain and cognition, vol. 1, no. 1, pp. 23–35, 1982.

46. D. Maurer, “Neonatal synesthesia: Implications for the processing of speech and faces,” in Developmental neurocognition: Speech and face processing in the first year of life, Springer, 1993, pp. 109–124.

47. G. Martino and L. E. Marks, “Synesthesia: Strong and weak,” Current Directions in Psychological Science, vol. 10, no. 2, pp. 61–65, 2001.

48. R. E. Cytowic, “Synesthesia: Phenomenology and neuropsychology,” Psyche, vol. 2, no. 10, pp. 2–10, 1995.

49. D. M. Eagleman, A. D. Kagan, S. S. Nelson, D. Sagaram, and A. K. Sarma, “A standardized test battery for the study of synesthesia,” Journal of neuroscience methods, vol. 159, no. 1, pp. 139–145, 2007.

50. R. E. Cytowic and D. Eagleman, Wednesday is indigo blue: Discovering the brain of synesthesia. MIT Press, 2009.

51. F. Spector and D. Maurer, “Synesthesia: a new approach to understanding the development of perception.,” Developmental psychology, vol. 45, no. 1, p. 175, 2009.

52. S. Shoemaker, “Self-knowledge and self-identity,” 1963.

53. S. Shoemaker, “Identity, cause, and mind: Philosophical essays,” 2004.

54. J. Lacan, “The Seminar of Jacques Lacan Book II: The Ego in Freud’s Theory and in the Technique of Psychoanalysis 1954–55,” Trans. Sylvana Tomaselli. Ed. Jacques-Alain Miller. New York: Norton, 1988.

55. J. Schaffer, “Monism: The priority of the whole,” Philosophical Review, vol. 119, no. 1, pp. 31–76, 2010.

56. D. Braddon-Mitchell, “The philosophy of mind and cognition,” 2007.

57. W. G. Lycan, “Inverted spectrum,” 1973.

58. D. R. Hilbert and M. E. Kalderon, “Color and the inverted spectrum,” Color perception: Philosophical, psychological, artistic, and computational perspectives, pp. 187–214, 2000.

59. M. Tye, “Qualia, content, and the inverted spectrum,” Noûs, pp. 159–183, 1994.

60. G. Harman, “The intrinsic quality of experience,” Philosophical perspectives, pp. 31–52, 1990.

61. J. Broackes, “Black and white and the inverted spectrum,” The Philosophical Quarterly, vol. 57, no. 227, pp. 161–175, 2007.

62. D. Cole, “Functionalism and inverted spectra,” in Epistemology and Cognition, Springer, 1991, pp. 85–100.

63. R. Villard and Z. Levay, “Creating Hubble’s Technicolor Universe,” Sky and Telescope, vol. 104, no. 3, p. 28, 2002.

64. A. Ventura, “Pretty Pictures: The Use of False Color in Images of Deep Space,” 2013.

65. R. Peierls, “Wolfgang Ernst Pauli. 1900-1958,” Biographical Memoirs of Fellows of the Royal Society, vol. 5, pp. 175–192, 1960.

66. C. H. Bennett, “Universal computation and physical dynamics,” Physica D: Nonlinear Phenomena, vol. 86, no. 1, pp. 268–273, 1995.

67. W. D. Hillis, The pattern on the stone: the simple ideas that make computers work. Basic Books, 2015.

68. M. Wilson, “Six views of embodied cognition,” Psychonomic bulletin & review, vol. 9, no. 4, pp. 625–636, 2002.

69. M. L. Anderson, “Embodied cognition: A field guide,” Artificial intelligence, vol. 149, no. 1, pp. 91–130, 2003.

70. L. Shapiro, Embodied cognition. Routledge, 2010.

71. J. C. Turner, M. A. Hogg, P. J. Oakes, S. D. Reicher, and M. S. Wetherell, “Rediscovering the social group: A self-categorization theory.,” Contemporary Sociology, 1987.

72. C. Calhoun, Social theory and the politics of identity. Blackwell, 1994.

73. B. E. Ashforth and F. Mael, “Social identity theory and the organization,” Academy of management review, vol. 14, no. 1, pp. 20–39, 1989.

74. J. C. Turner, “Towards a cognitive redefinition of the social group,” Social identity and intergroup relations, pp. 15–40, 1982.

75. R. Centers, “The psychology of social classes: a study of class consciousness.,” 1949.

76. W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” The bulletin of mathematical biophysics, vol. 5, no. 4, pp. 115–133, 1943.

77. L. N. De Castro and F. J. Von Zuben, Recent developments in biologically inspired computing. Igi Global, 2005.

78. L. Bochereau and P. Bourgine, “Extraction of semantic features and logical rules from a multilayer neural network,” in Proceedings of the International Joint Conference on Neural Networks, 1990, pp. 579–582.

79. Y. Hayashi, “A neural expert system with automated extraction of fuzzy if-then rules,” in Advances in neural information processing systems, 1991, pp. 578–584.

80. G. G. Towell, “Symbolic knowledge and neural networks: Insertion, refinement and extraction,” 1992.

81. G. Towell and J. W. Shavlik, “Interpretation of Artificial Neural Networks:...,” 1992.

82. M. W. Craven and J. W. Shavlik, “Understanding neural networks via rule extraction and pruning,” in Proceedings of the 1993 Connectionist Models Summer School, 1994, p. 184.

83. D. A. Hall and D. R. Moore, “Auditory neuroscience: The salience of looming sounds,” Current Biology, vol. 13, no. 3, pp. R91–R93, 2003.

84. W. Schneider and R. M. Shiffrin, “Controlled and automatic human information processing: I. Detection, search, and attention.,” Psychological review, vol. 84, no. 1, p. 1, 1977.

85. X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on, 2007, pp. 1–8.

86. M. I. Posner and C. R. Snyder, “Attention and Cognitive Control,” Cognitive psychology: Key readings, p. 205, 2004.

87. D. Kahneman, Attention and effort. Citeseer, 1973.

88. R. D. Hare, The Hare Psychopathy Checklist-Revised: PLC-R. MHS, Multi-Health Systems, 1999.

89. J. M. Fischer, “Compatibilism,” 2007.

90. J. A. Gray, J. A. Gray, and J. A. Gray, "Consciousness: Creeping up on the hard problem," 2004.

91. T. Horgan, "Functionalism, qualia, and the inverted spectrum," Philosophy and Phenomenological Research, pp. 453–469, 1984.

92. S. Shoemaker, "The inverted spectrum," The Journal of Philosophy, pp. 357–381, 1982.

93. C. Klein, "What pain asymbolia really shows," Published online at philpapers. org/rec/KLEWPA, 2011.

▲ Return to Top